DOCUMENT RESUME

ED 452 213 TM 032 498

AUTHOR	Shen, Linjun
TITLE	A Comparison of Angoff and Rasch Model Based Item Map Methods in Standard Setting.
PUB DATE	2001-04-00
NOTE	9p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Comparative Analysis; *Licensing Examinations (Professions); *Physicians; *Standards; *Test Items
IDENTIFIERS	*Angoff Methods; Mapping; *Rasch Model; Standard Setting

ABSTRACT

Two standard setting methods, the Angoff method and the Rasch model based Item Map approach, were compared for setting a standard for a high-stakes medical licensure examination, the last examination of a three-examination series of a national medical licensing examination. The standard setting committee consisted of 23 physicians who were involved in postgraduate medical education. For the Angoff method, the 800 test items were divided into 2 sets, and each member reviewed about 480 items. For the Item Map method, 13 maps were constructed based on the combination of specialty and dimension. Each map had five to seven items ordered by the item difficulties calibrated from a previous examination, and judges were asked to draw a line on a map to indicate that items below the line were those candidates should master in order to start practicing medicine and items above the line were those candidates could master later. The values of the lines were the difficulty values, in logits of the items immediately below the lines. The standard set by the Angoff method was substantially more severe than the standard s et by the Item Map method, but the latter better reflected the desired content criterion of minimum competence. The result was more practical, and the standard set was consistent with historical data. The Rasch model method was, in this case, a truly criterion-referenced method. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document

ED 452 213

A Comparison of Angoff and Rasch Model Based Item Map Methods in Standard Setting

Linjun Shen

National Board of Osteopathic Medical Examiners

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
L. Shen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM032498

Paper accepted for presentation at the Annual Meeting of the American Educational Research Association

Seattle, April 2001

BEST COPY AVAILABLE

The Angoff standard setting approach has been the most widely used method in medical licensure and certification. The question asked by the Angoff method is, "What is the probability that a minimally competent candidate would answer this question correctly?" The estimated performance standard for a judge is determined by summing the estimates for all items in an exam. The resulting average "Angoff rating" over judges is used as the performance standard for the examination. There are varieties of modifications of this approach. Some provide judges real performance data from the exam for which the standard is to be set to adjust their ratings[1]. Others use performance data on the previous exams to train judges[2]

Rasch-model based Item Map approach for standard setting has also been applied in medical licensure and certification[3,4], though not as commonly as the Angoff method. The assumption of this approach is that the item difficulty order calibrated by Rasch model reflects the order of the content complexity. What judges need to do is to decide the minimal competence level within the hierarchy of the content.

Both methods claim themselves valid, criterion-referenced, and efficient standard setting approaches. The purpose of this paper was to compare the two methods in setting a standard for a high-stake medical licensure examination. The focuses of the comparison were: the content definitions of the standards, practicality of the resulting standards, and efficiency of the methods.

Methods

Examination

The examination the two methods set standard for was the last exam of a three exam series of a national medical licensing exam program. A passage of this exam provides an eligibility for unrestricted medical license. Candidates of the exam were medical graduates with at least 6 months of postgraduate training. The exam had 800 multiple-choice questions constructed according to a two dimensional blueprint.

Standard setting committee

The standard setting committee comprised of a total of 23 physicians who covered all major clinical specialties and had a good geographic representation. All members were substantially involved in postgraduate medical education.

The Angoff method

For the Angoff method, 800 items were divided into two sets. Half of the Committee reviewed one set, the other another set. In addition, there were 80 common ittems used to assess the judge consistency. Overall, each member reviewed about 480 items. The committee spent an hour to discuss and establish the concept of "borderline examinees". Afterward, it went through two practices with performance data available. Eight hours were allocated for the Angoff rating of 480 items.

The Item-Map method

Thirteen maps were constructed based on the combination of specialty and dimension 2 categories. Figure 1 demonstrates the specifics of the maps. The emphases of the 13 maps were Management and primary care specialties. Each map had 5-7 items ordered by the item difficulties calibrated from a previous exam. Full item texts of the items in each map were provided to judges also. Judges were asked to draw a line in a map to indicate that items below the line were the items candidates should master in order to start practicing medicine and items above the line were the items the candidates can master later. The values of the lines were the difficulty values (in logit) of the items immediately below the lines. The final standard created by the Map approach was the mean of all the lines over all the judges. Judges were given three practice maps to get familiar with the methods.

In addition to the two methods, Hofstee method[4] was also embedded in Angoff method. After every 90 items, a short questionnaire was given to the judges asking the range of minimum and maximum acceptable passing percentages and percentage of correct based on the 90 items only. Those four mean values were used to plot two points. One point represented the minimum acceptable performance standard and the maximum acceptable failure rate. The second was defined by the maximum acceptable performance standard and minimum acceptable failure rate. The intersection of the line segment with the cumulative frequency curve of the student actual performance on the exam defined the standard.

Results

The Angoff standard

Correlations between the first and second panels was .70. The inter-rater reliability of 23 judges on the 80 common items was .80. The inter-rater reliability of 12 even numbered judges was .76. The inter-rater reliability of 11 odd numbered judges was .75.

The mean rating of Even numbered group was 63.52% and for Odd numbered group was 64.30%. The combined mean percentage correct was 63.9%. This would fail 21.5% of the candidates. The 95% confidence interval of this standard was 63.08% to 64.82%, equivalent to the failing rates of 17.4% to 26.1%.

The Map approach

The mean of the minimum competence level of the 13 maps was .77 logits. This was converted to the maximum failing of .76 corresponding to a failing rate of 12.5%. The median of the minimum competence level of the 13 maps was .74. This was converted to the maximum failing of .73 or a failing rate of 11.1%.

Hofstee method

The overall judgment of the minimum percentage of items candidates should get right was 53.27%, the maximum was 69.72%. The overall judgment of minimum failing rate was 8.96%, the maximum was 17.89%.

These four estimates, interrelated to the performance curve, yielded 62% as the minimum percentage of items candidates should get correct. This was equivalent to a failing rate of 11.9%.

Previous failure rates

The failure rates of the same exam in the past five years were: 8% in 1999, 7% in 1998 and 1997, 6% in 1996, and 9% in 1995.

Figure 2 compares the standards proposed by Angoff and Item Map methods with Hofstee standard and historical failure rates as references.

Discussions

Obviously, the standard set by Angoff method was substantially severe than the standard by Item Map method as well as the Hofstee standard and all the previous failure rates. To evaluate which standard specified the minimum competence most meaningfully and reasonably, a content analysis of the minimum passing levels implied by those standards was inevitable. The standard from Angoff method was simply a decision of percentage correct. There was no explicit content definition of the standard. On the other hand, the standard from the Item Map method was directly derived from separating items candidates should have mastered now from those to be mastered later. So, the minimum level of competence was clearly defined by those items immediately above the standard lines. In order to compare the substantive meaning of the two standards, the standard from Angoff method was first converted into a Rasch logit of .87. Then item contents at the logit .87 level were compared with items at .73-.76 level (the Rasch standard). The comparisons revealed that the Angoff standard often demanded more speciality-oriented knowledge, on the other hand, the Item Map standard often demanded more fundamental knowledge which were definitely required to practice medicine without supervision.

In comparing the standards with the historical failure rates, Item Map standard was progressively higher but not as dramatic as the Angoff standard. Hofstee standard was more normative than criterion-referenced. Nevertheless, it was another way for reality checking. The Map standard and Hofstee standard agreed well, while the Angoff standard was substantially different from the Hofstee standard.

In summary, the standard set by the Item Map approach in comparison with the standard set by the Angoff method better reflected the desired content criterion of minimum competence, the consequence was more practical, and was consistent with the historical data.

The difficulty of the Angoff method was that it could not clearly define the content criterion of the minimum competence. The method entirely depended on the concept of "borderline examinees". In other words, the Angoff method used "borderline examinees" as the proxy for the minimum competency. Since "borderline examinees" could never been operationally defined, the outcomes of the method were totally results of judges' educational guessing.

Some variations of the Angoff method provide real performance data to judges for them to adjust their rating. If such modifications were used in this study, the standard might have been less severe. However, that type of practice would make the Angoff method even less criterion-referenced and more normal-referenced.

Efficiency of the standard setting process is another dimension of the methods. In this study, 12 hours were scheduled for the Angoff process. It actually took 8 hours. For Map method, 3.5 hours were planned and 3 hours were actually used. Rating 480 items was tedious and tiring. Reviewing 13 maps was relatively straightforward and rather enjoyable. From the efficiency perspective, the Map approach was more favorable than the Angoff method.

Overall, the Rasch model based item map approach worked better in this particular study. It was truly a criterion-referenced method. The judges' decisions were based on direct content review. The final decision had explicit specifications in terms of content. The outcomes were meaningful and reasonable. The process was more efficient. One unique requirement by the Map method is that it requires some previously used items embedded in the exam while the Angoff method, theoretically, does not have this requirement. In practice, this requirement may limit the application of the Map method.

References

1. Shepard, L.A. (1984). Setting performance standards. In R.A. Berk (Ed.) *A Guide to criterion-referenced test construction.* Baltimore: John Hopkins Press.

2. Swanson, D.B., Dillon, G.F., and Ross, L.P. Setting content-based standards for National Board exams: initial research for the comprehensive Part I Examination. Academic Medicine: 65, S17-18.

3. Grosse M.E. and Wright, B.D. Setting, evaluating, and maintaining certification standards with the Rasch Model. Evaluation & the Health Professions, 9: 267-285.

4. Stone, G.E. A standard vision. Popular Measurement: journal of the Institute for Objective Measurement: 3,40-41.

Figure 1

Content Specifications of Maps Constructed for Standard Setting

	Surgery	OB/Gyn	Psych	FM	Peds	Med	EM	OPP
Health Promotion				X		X		
History and Physical		X			X			
Diagnos. Technology					X		X	
Management	X	X	X	X				X
Disease Mechanisms								X
Health Care Delivery						X		

Figure 2
Comparison of the Angoff and Map Standards

Percentage Failure	Standards of Previous Exams	Standards Proposed by	Failure Rate
23			
22			
21		Angoff	21%
20			
19			
18			
17			
16			
15			
14			
13			
12		Hofstee	12%
11		MAP Median	11%
10		MAP Mean	10%
9	95/1		
8	99/1		
7	98/1 97/1		
6	96/1		
5			
4			
3			
2			
1			
0			